

Docket No.: PC-0034 US

REMARKS

Claims 1-20 were originally filed and were subject to a Restriction Requirement. Applicants affirm election, with traverse, of original claims 1-6, corresponding to the invention of Group I.

Justification for the amendments is as follows. The specification has been amended to delete reference to certain web sites recited in the application. No new matter is added by any of these amendments.

Drawings

The Examiner noted that the Brief Description of the Figures and Table on page 5 of the specification contains a referral to Tables 1 and 2. However, the instant specification is missing the Tables. Clarification is required. Applicants file record for this case shows that Tables 1 and 2 were indeed submitted with the application as originally filed as pages 38 and 39. Enclosed is a copy of the Return Postcard clearly showing the referenced Tables 1 and 2 as pages 38 and 39 of the specification. For the Examiner's convenience, a copy of Tables 1 and 2, as originally filed, is also attached to this response.

Specification

The disclosure is objected to because it contains an embedded hyperlink and/or other form of browser-executable code, see p. 28, line 10 and p.29, line 21, for example. Applicant is required to delete the embedded hyperlink and/or other form of browser-executable code. See MPEP § 608.01.

Applicants submit that the MPEP states at § 608.01 that this policy is based on the principle that "USPTO policy does not permit the USPTO to link to any commercial sites since the USPTO exercises no control over the organization, views or accuracy of the information contained on those outside sites (underline added). Section 608.01 goes on to state that "where hyperlinks and/or other forms of browser-executable codes are a part of the applicant's invention and it is necessary to have them included in the patent application in order to comply with the requirements of 35 U.S.C. 112, first paragraph, and applicant does not intend to have these hyperlinks as active links, examiners should not object to these hyperlinks. The Office will disable these hyperlinks when preparing the text to be

Docket No.: PC-0034 US

loaded onto the USPTO web database (underline added). Applicants point out that the cited website is a non-commercial, government web site which should not be subject to the requirements of MPEP § 608.01. However, this citation, as well as a second at page 33, line 26 have been deleted. Withdrawal of the objection is therefore requested.

35 U.S.C § 101, Rejection of Claims 1-6

The Examiner has rejected claims 1-6 under 35 U.S.C. § 101 because it is drawn to an invention with no apparent or disclosed specific and substantial credible utility. The instant application has provided a description of an isolated DNA encoding a protein and the protein encoded thereby. The instant application does not disclose the biological role of this protein or its significance. The Examiner stated that it is clear from the instant application that the protein described is what is termed an "orphan protein" in the art and that there is little doubt that after complete characterization this DNA and encoded protein may be found to have a specific and substantial utility, however, that this further characterization is part of the act of invention. See *Brenner V. Manson* 148 USPQ 689 (Sus. Ct, 1966).

The Examiner reiterated Applicants characterization of the human TIMM8b to which the instant protein, TRP, bears sequence homology to (85% amino acid identity) and its association with various neurodegenerative and neuromuscular diseases involving defects in oxidative phosphorylation. The Examiner, however, cited various publications regarding an alleged uncertainty in the art in predicting protein function based on structure (Skolnick et al. (2000) and Bork et al. (1998)) and concluded that the function of TRP could not unequivocally be extrapolated from its structural characteristics (underline added). The Examiner concluded that the instant specification fails to provide any evidence or sound scientific reasoning that would support a conclusion that the instant nucleic acid or encoded protein is associated with any diseases or disorders.

Applicants disagree that the claimed invention is not supported by either a well-established utility or a specific and substantial asserted utility. The claimed invention is in fact supported by both a well established utility and a specific and substantial, asserted utility.

TRP is supported by a well established utility based on its structural and functional identity with TIMM8b, disclosed in the specification as a mitochondrial protein involved in neurodegenerative

Docket No.: PC-0034 US

disorders, such as Mohr-Tranebjaerg syndrome. See BACKGROUND, pp. 1-2 and Paschen *et al* (2000). The identification of TRP as a TIMM8b protein is based on a high level of sequence identity to TIMM8b. In addition to an overall sequence identity of 85% with TIMM8b, The sequence alignment presented in Figure 2 clearly shows that TRP is 100% identical with TIMM8b over 82% if the sequence of TRP, differing only by a 15 amino acid insert at the N-terminal end of the molecule. It is well known in the art that such N-terminal sequences are most likely signal peptides related to protein secretion or subcellular localization rather than to altering function. In fact, as described in the specification at p. 9, line 30-31, TRP retains (100% identical) the $CX_3CX_{14}CX_3C$ motif of mitochondrial import proteins and which is characteristic of DDP/TIM family proteins. Thus there is a substantial likelihood that TRP is functionally as well as structurally related to other DDP/TIM family proteins, such as TIMM8b. In addition, It is well-known that the probability that two unrelated polypeptides share more than 40% sequence homology over 70 amino acid residues is exceedingly small. Brenner *et al.*, *Proc. Natl. Acad. Sci.* 95:6073-78 (1998). Given homology in substantial excess of 40% over many more than 70 amino acid residues, including the conservation of functional motifs, the probability that the polypeptide encoded for by the claimed polynucleotide is related to TIMM8b is, accordingly, very high. The Examiner must accept the applicants' demonstration that the homology between the polypeptide encoded for by the claimed invention and TIMM8b demonstrates utility by a reasonable probability unless the Examiner can demonstrate through evidence or sound scientific reasoning that a person of ordinary skill in the art would doubt utility. See *In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Examiner has not provided sufficient evidence or sound scientific reasoning to the contrary. While the Examiner has cited literature identifying some of the difficulties that may be involved in predicting protein function, none suggests that functional homology cannot be inferred by a reasonable probability in this case. See Skolnick *et al.* and Bork *et al.*, Office Action, p. 5. Most important, none contradicts Brenner's basic rule that sequence homology in excess of 40% over 70 or more amino acid residues yields a high probability of functional homology as well. Nor do they contradict the significance of the $CX_3CX_{14}CX_3C$ motif retained in TRP. At most, these articles individually and together stand for the proposition that it is difficult to

Docket No.: PC-0034 US

make predictions about function with certainty. The standard applicable in this case is not, however, proof to certainty (or unequivocal proof), but rather proof to reasonable probability.

In addition, the claimed polynucleotide is also supported by a specific and substantial asserted utility that is independent of any knowledge of the encoded protein. This utility is disclosed in the specification at p. 17, lines 4-7 where it is stated "The cDNAs, fragments, oligonucleotides, complementary RNA and DNA molecules, and PNAs and may be used to detect and quantify differential gene expression for diagnosis of a disorder. Disorders associated with differential expression include cancer, particularly breast cancer, ovarian cancer, and kidney cancer--". This asserted utility of polynucleotides encoding TRP in the diagnosis of breast, ovarian, and kidney cancer is supported in the specification at p. 9, in the paragraph beginning at line 9 and Table 2, where it is shown that SEQ ID NO:2 shows overexpression in a breast tumor library (BRSTTUT14) compared with microscopically normal breast tissue from the same donor (BRSTNOT14) in which no expression of the transcript was detectable. Similarly, SEQ ID NO:2 was overexpressed in two kidney tumor libraries (KIDNTUT15 and KIDNTUT14) compared to libraries (KIDNNOT19 and KIDNNOT20) from matched (m) microscopically normal tissue from the same donors. SEQ ID NO:2 was also overexpressed in two ovarian tumor libraries (OVRTUT02 and OVRTUT03). Thus the claimed polynucleotide is useful in the detection and diagnosis of breast, ovarian, and kidney cancers independent of any knowledge of the polypeptide encoded by the polynucleotide.

For all of the above reasons, Applicants believe the claimed invention is well supported by both a well established utility as a functional homolog of TIMM8b in the diagnosis of neurodegenerative disorders, such as Mohr-Tranebjaerg syndrome, as well as a specific and substantial asserted utility in the diagnosis of breast, ovarian, and kidney cancers. Withdrawal of the rejection of claims 1-6 under 35 U.S.C. § 101 is therefore respectfully requested.

35 U.S.C. § 112, First Paragraph, Rejection of Claims 1-6

The Examiner has also rejected claims 1-6 under 35 U.S.C § 112, first paragraph, specifically, since the claimed invention is not supported by either a clear asserted utility or a well established utility for the reasons set forth above, one skilled in the art would clearly not know how to use the invention.

Docket No.: PC-0034 US

The rejection of claims 1-6 under 35 U.S.C. § 101 for lack of utility has been addressed above. Applicants therefore submit that, since the claimed invention is supported by both a well established utility, as well as a specific and substantial asserted utility, one skilled in the art would clearly know how to use the claimed invention. Withdrawal of the rejection of claims 1-6 under 35 U.S.C § 112, first paragraph is therefore requested.

35 U.S.C. 112, Second Paragraph, Rejection of Claims 2 and 6

The Examiner has rejected claims 2 and 6 under 35 U.S.C. § 112, second paragraph as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention. In particular, the Examiner stated that claim 2 is vague and ambiguous for recitation of "a fragment" of SEQ ID NO:5. Although according to the definition presented in the specification, "[f]ragment refers to a chains of consecutive nucleotides from about 50 to about 4000 base pairs in length" a general meaning of "a fragment" is "a part of something". The Examiner stated that it is not clear how a sequence of SEQ ID NO:2, which is 455 nucleotides long can have a fragment that is a sequence of SEQ ID NO:5, which is 598 nucleotides long. Clarification is required.

EST fragments of full-length consensus sequences, such as the instant SEQ ID NO:2, sometimes contain untranslated regions (UTR) that do not align with the full length sequence. As the attached alignment (Exhibit A) shows SEQ ID NO:5 (Incyte Clone 1661626F6) aligns over 321 base pairs of its sequence with the majority of the open reading frame of SEQ ID NO:2 and is clearly therefore a fragment of SEQ ID NO:2 from which the consensus sequence is derived.

The Examiner stated that claim 6 is vague and indefinite for recitation of "a protein", which is produced by the host cell of claim 5. It is not clear which protein produced by a host cell is intended by the claim, because the claim depends from claim 1, which encompasses both coding and non-coding sequences of a nucleic acid. The Examiner also stated that claim 6 is indefinite and ambiguous for missing a critical relationship because the claim is not limited to the host cell of claim 5 or the vector of claim 4, but depends from claim 1, which encompasses a nucleic acid sequence encoding a protein having the amino acid sequence of SEQ ID NO:1 and also a nucleotide sequence that is completely complementary to the nucleic acid encoding SEQ ID NO:1.

Docket No.: PC-0034 US

Both claims 4 and 5 indeed do depend from claim 1 and are therefore subject to the limitations of the polynucleotides recited in claim 1, e.g., a polynucleotide encoding SEQ ID NO:1 or its complete complement. Since the complementary sequence to a polynucleotide encoding an open reading can, itself, encode a protein, it is customary to include it any expression system for expressing a protein, such as that recited in claim 6. With these explanations, Applicants believe claims 2 and 6 are clear and definite, and respectfully request withdrawal of the rejection of these claims under 35 U.S.C. § 112, second paragraph.

Docket No.: PC-0034 US

CONCLUSION

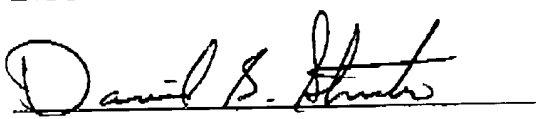
In light of the above amendments and remarks, Applicants submit that the present application is fully in condition for allowance, and request that the Examiner withdraw the outstanding rejections. Early notice to that effect is earnestly solicited. Applicants further request that, upon allowance of claim 1, claims 7-12 be rejoined and examined as methods of use of the polynucleotides of claim 1 that depend from and are of the same scope as claim 1 in accordance with *Ochiai and Brouwer*. See MPEP § 821.04 and the Commissioner's Notice in the Official Gazette of March 26, 1996.

If the Examiner contemplates other action, or if a telephone conference would expedite allowance of the claims, Applicants invite the Examiner to contact Applicants' agent of Record, below.

Applicants believe that no fee is due with this communication. However, if the USPTO determines that a fee is due, the Commissioner is hereby authorized to charge Deposit Account No. 09-0108.

Respectfully submitted,

INCYTE GENOMICS, INC.

Date: October 10, 2002

David G. Streeter, Ph.D.

Reg. No. 43,168

Direct Dial Telephone: (650) 845-5741

3160 Porter Drive
Palo Alto, California 94304
Phone: (650) 855-0555
Fax: (650) 849-8886

Docket No.: PC-0034 US

VERSION WITH MARKINGS TO SHOW CHANGES MADE**IN THE SPECIFICATION**

Paragraph beginning at line 10 of page 28 has been amended as follows:

The BLAST software suite (NCBI, Bethesda MD[; <http://www.ncbi.nlm.nih.gov/gorf/bl2.html>]), includes various sequence analysis programs including "blastn" that is used to align nucleotide sequences and BLAST2 that is used for direct pairwise comparison of either nucleotide or amino acid sequences. BLAST programs are commonly used with gap and other parameters set to default settings, e.g.: Matrix: BLOSUM62; Reward for match: 1; Penalty for mismatch: -2; Open Gap: 5 and Extension Gap: 2 penalties; Gap x drop-off: 50; Expect: 10; Word Size: 11; and Filter: on. Identity is measured over the entire length of a sequence. Brenner et al. (1998; Proc Natl Acad Sci 95:6073-6078, incorporated herein by reference) analyzed BLAST for its ability to identify structural homologs by sequence identity and found 30% identity is a reliable threshold for sequence alignments of at least 150 residues and 40%, for alignments of at least 70 residues.

Paragraph beginning at line 15 of page 29 has been amended as follows:

Following assembly, templates were subjected to BLAST, motif, and other functional analyses and categorized in protein hierarchies using methods described in USSN 08/812,290 and USSN 08/811,758, both filed March 6, 1997; in USSN 08/947,845, filed October 9, 1997; and in USSN 09/034,807, filed March 4, 1998. Then templates were analyzed by translating each template in all three forward reading frames and searching each translation against the PFAM database of hidden Markov model-based protein families and domains using the HMMER software package (Washington University School of Medicine, St. Louis MO[; <http://pfam.wustl.edu/>]). The cDNA was further analyzed using MACDNASIS PRO software (Hitachi Software Engineering), and LASERGENE software (DNASTAR) and queried against public databases such as the GenBank rodent, mammalian, vertebrate, prokaryote, and eukaryote databases, SwissProt, BLOCKS, PRINTS, PFAM, and Prosite.

Oct 10, 2002

5:36PM

INCYTE LEGAL DEPT

No. 8808 P. 12

Express Mail No.: EL 743 381 725 US

Mailed: February 8, 2001
Docket No.: PC-0034 US

COMMISSIONER FOR PATENTS
BOX PATENT APPLICATION
WASHINGTON, D.C. 20231

Inventors: Jennifer L. Hillman
Title: TIMM8b-RELATED PROTEIN
Filing Date: Herewith

Enclosed:

- ☒ Return Receipt Postcard
- ☒ Transmittal for Patent Application (1 page, in duplicate)
- ☒ Submission of Sequence Listing (1 page)
- ☒ One Computer-readable Diskette
- ☒ 37 Pages of Specification (1-37);
- ☒ 2 Pages of Tables (Table 1-2) (38-39);
- ☒ 2 Pages of Claims (40-41);
- ☒ 1 Page of Abstract (42);
- ☒ 3 Sheets of Figures (1A, 1B, and 2)
- ☒ 5 Pages of Sequence Listing (1-5); and
- ☒ 3 Pages - Unexecuted Declaration and Power of Attorney.

DGS/ks

Express Mail No.: EL 743 381 725 US

Mailed: February 8, 2001
Docket No.: PC-0034 US

COMMISSIONER FOR PATENTS
BOX PATENT APPLICATION
WASHINGTON, D.C. 20231

Inventors: Jennifer L. Hillman
Title: TIMM8b-RELATED PROTEIN
Filing Date: Herewith

Enclosed:

- ☒ Return Receipt Postcard
- ☒ Transmittal for Patent Application (1 page, in duplicate)
- ☒ Submission of Sequence Listing (1 page)
- ☒ One Computer-readable Diskette
- ☒ 37 Pages of Specification (1-37);
- ☒ 2 Pages of Tables (Table 1-2) (38-39);
- ☒ 2 Pages of Claims (40-41);
- ☒ 1 Page of Abstract (42);
- ☒ 3 Sheets of Figures (1A, 1B, and 2)
- ☒ 5 Pages of Sequence Listing (1-5); and
- ☒ 3 Pages - Unexecuted Declaration and Power of Attorney.

DGS/ks

J1036 U.S. PTO

09/781117



02/08/01

Tissue Category	Clone Count	Found in	Abs Abund	Pct Abund
Cardiovascular System	256190	17/68	32	0.0120
Connective Tissue	144645	10/47	11	0.0076
Digestive System	501101	23/148	34	0.0068
Embryonic Structures	106713	5/21	17	0.0159
Endocrine System	225386	14/53	20	0.0089
Exocrine Glands	254635	15/64	21	0.0082
Reproductive, Female	427284	20/106	28	0.0066
Reproductive, Male	448207	24/114	32	0.0071
Germ Cells	38282	3/5	5	0.0131
Hemic and Immune System	680277	35/159	55	0.0081
Liver	109378	7/35	17	0.0155
Musculoskeletal System	159280	9/47	11	0.0069
Nervous System	955753	55/198	76	0.0080
Pancreas	110207	1/24	1	0.0009
Respiratory System	390086	23/93	33	0.0085
Sense Organs	19256	0/8	0	0.0000
Skin	72292	3/15	4	0.0055
Stomatognathic System	12923	1/10	3	0.0232
Unclassified/Mixed	120926	3/13	13	0.0108
Urinary Tract	279062	10/64	19	0.0068
Totals	5321883	278/1292	432	0.0081

TABLE 1

Found in:

<u>Library ID</u>	<u>Clone Count</u>	<u>Library Description</u>	<u>Abs Abund</u>	<u>Pct Abund</u>
BRSTTUT14	3960	breast tumor, adenocA, 62F, m/BRSTNOT14	2	0.0505
OVARTUP02	3158	ovary tumor, serous papillary adenocA, F, 3' CGAP	2	0.0633
OVARTUT03	4249	ovary tumor, seroanaplastic CA, 52F	2	0.0471
KIDNTUT15	3954	kidney tumor, renal cell CA, 65M m/KIDNNOT19	2	0.0506
KIDNTUT14	3872	kidney tumor, renal cell CA, 43M, m/KIDNNOT20	1	0.0258

Not found in:

<u>Library ID</u>	<u>Clone Count</u>	<u>Library Description</u>
BRSTNOT14	3800	breast, mw/ductal adenocA, CA in situ, 62F, m/BRSTTUT14
KIDNNOT19	6963	kidney, mw/renal cell CA, 65M, m/KIDNTUT15
KIDNNOT20	3718	kidney, mw/renal cell CA, 43M, m/KIDNTUT14

TABLE 2

Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER^{1,2}, CYRUS CHOTHIA³, AND TIM J. P. HUBBARD⁴

¹MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and ²Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)

ABSTRACT Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T., & Chothia, C. (1993) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA (k_{up} = 1), and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

Previous Assessments of Sequence Comparison. Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith-Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (k_{up} = 2) or greater effectiveness (k_{up} = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/956073-06\$2.00/0
PNAS is available online at <http://www.pnas.org>

Abbreviation: EPO, errors per query.

¹Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126

²To whom reprints requests should be addressed. e-mail: brenner@hycr.stanford.edu.

superfamilies. Pearson found that modern matrices and "in-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith-Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18-20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

A Database for Testing Homology Detection. Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it

is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB40D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-S) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or =0.5% of the total 1,749,006 ordered pairs. In PDB40D-S, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from <http://ssa.stanford.edu/ssa/>, and databases derived from the current version of SCOP may be found at <http://scop.mrc-lmb.cam.ac.uk/scop/>.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB40D-S (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

Assessment Data and Procedure. Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0176 (3), which provided FASTA and the SSEARCH implementation of Smith-Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties =12/-1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

The "Coverage Vs. Error" Plot. To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

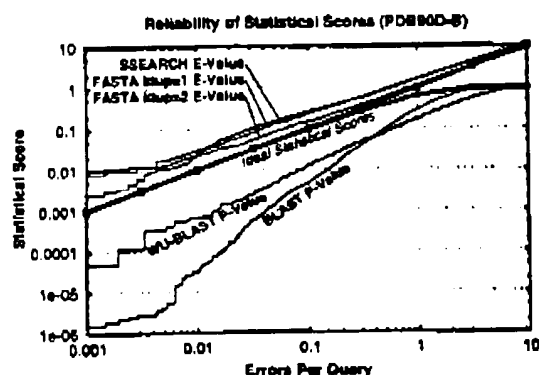


FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB90D-B were similar to those for PDB90D-A despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

The Performance of Scoring Schemes. All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith-Waterman" score, which is the measure optimized by the Smith-Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure

related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

Sequence Identity. Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

Raw Scores. Smith-Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

Statistical Scores. Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

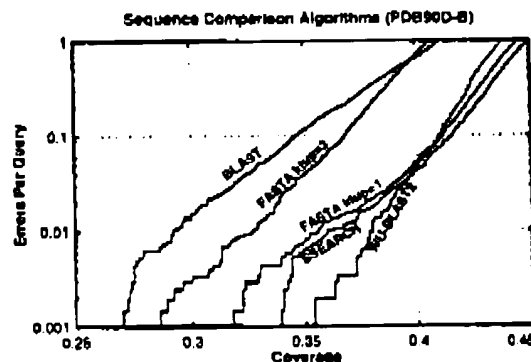
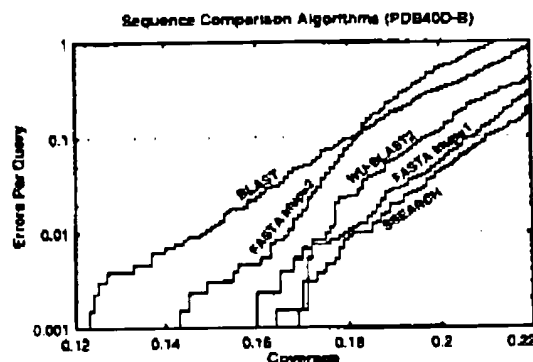


FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (A) PDB90D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA klup = 1 and WU-BLAST2 are almost as good. (B) PDB90D-A database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA klup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

Overall Detection of Homologs and Comparison of Algorithms. The results in Fig. 5A and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB400-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB400-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5B). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA ktup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

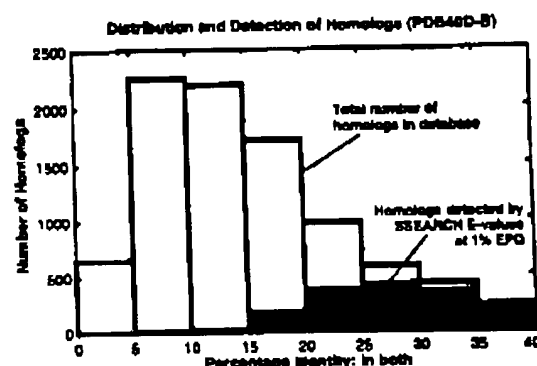


Fig. 6. Distribution and detection of homologs in PDB400-B. Bars show the distribution of homologous pairs PDB400-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB400-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pairwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (i) using a large current database in which the protein sequences have been complexity masked and (ii) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB400-B

Method	Relative Time*	1% EPQ Cutoff	Coverage at 1% EPQ
SSEARCH % identity: within alignment	25.5	>70%	<0.1
SSEARCH % identity: within both	25.5	34%	3.0
SSEARCH % identity: HSP-scaled	25.5	35% (HSP + 9.8)	4.0
SSEARCH Smith-Waterman raw scores	25.5	142	10.5
SSEARCH E-values	25.5	0.03	18.4
FASTA ktup = 1 E-values	3.9	0.03	17.9
FASTA ktup = 2 E-values	1.4	0.03	16.7
WU-BLAST2 P-values	1.1	0.003	17.5
BLAST P-values	1.0	0.00016	14.8

*Times are from large database searches with genome proteins.

extent of errors. Second, sSEARCH, WU-BLAST, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at <http://sss.stanford.edu/ssw/>.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

- Altshul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* 215, 403-410.
- Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460-480.
- Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444-2448.
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* 247, 536-540.
- Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* 266, 635-643.
- Pearson, W. R. (1991) *Genomics* 11, 635-650.
- Pearson, W. R. (1995) *Protein Sci.* 4, 1145-1160.
- Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195-197.
- George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* 266, 41-59.
- Vogt, G., Ezzold, T. & Argos, P. (1995) *J. Mol. Biol.* 249, 816-831.
- Henikoff, S. & Henikoff, J. G. (1993) *Proteins* 17, 49-61.
- Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* 24, 21-25.
- Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* 24, 189-196.
- Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* 89, 10915-10919.
- Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Bio-medical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345-352.
- Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
- Sander, C. & Schneider, R. (1991) *Proteins* 9, 56-68.
- Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* 233, 716-738.
- Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* 1, 89-94.
- Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* 1, 77-78.
- Arratia, R., Gordon, L. & M. W. (1986) *Ann. Stat.* 14, 971-993.
- Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* 87, 2264-2268.
- Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* 90, 5873-5877.
- Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* 6, 119-129.
- Pearson, W. R. (1996) *Methods Enzymol.* 266, 227-258.
- Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* 12, 215-226.
- Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* 266, 554-571.
- Waterman, M. S. & Vingron, M. (1994) *Stat. Science* 9, 367-381.
- Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* 13, 669-678.
- Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Int'l. Union Crystallogr., Cambridge, UK), pp. 107-132.
- Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* 7, 369-376.
- Orengo, C., Michie, A., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. (1997) *Structure (London)* 5, 1093-1108.
- Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* 39, 561-577.
- Gribokov, M. & Robinson, N. L. (1996) *Comput. Chem.* 20, 25-33.
- Fitch, W. M. (1966) *J. Mol. Biol.* 16, 9-16.
- Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* 4, 1123-1127.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* 25, 3389-3402.
- Girling, R., Schmidt, W., Jr., Houston, T., Amma, E. & Huismann, T. (1979) *J. Mol. Biol.* 131, 417-433.
- Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* 32, 9906-9916.
- Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* 20, 374-376.

